White Paper Report

Report ID: 98474

Application Number: HD5030008

Project Director: Mark LeBlanc (mleblanc@wheatoncollege.edu)

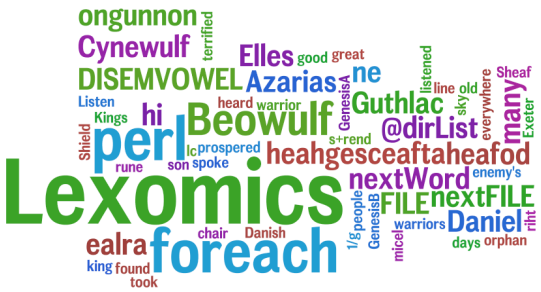
Institution: Wheaton College

Reporting Period: 7/1/2008-6/30/2010

Report Due: 9/30/2010

Date Submitted: 9/8/2010

Pattern Recognition through Computational Stylistics: Old English and Beyond



White Paper to the NEH Office of Digital Humanities
Level 2 Digital Humanities Start-Up Grant
HD 50300-08, 2008-2010

April 2010

Mark D. LeBlanc[1] (Project Director), Computer Science
Michael Drout (Co-PI), English
Michael Kahn (Co-PI), Statistics

Our project received Digital Humanities Level 2 Start-Up funding ($41,950) to support the design and implementation of a suite of computational tools, software, and statistical analyses to explore the Old English corpus. The work serves as a proof of concept for the larger deployment of corpus-independent tools. Outcomes include scalable, open-source software to facilitate the computation and organization of word frequencies across an entire corpus, a website facilitating software dissemination, and unique ways of viewing the details of user-defined (virtual) manuscripts of Anglo-Saxon poetry and prose. This research has significantly influenced the development of interdisciplinary course materials for our "connected" (interdisciplinary) digital humanities undergraduate courses in English, Statistics, and Computer Science. We have disseminated our experimental results in a prestigious, peer-reviewed journal (*The Journal of English and Germanic Philology*) as well as in presentations at national and international conferences. A small initial investment from the NEH enabled us to develop and disseminate our software and tools to a wide range of scholars and to begin collaborations with other researchers both nationally and internationally.

[1] Contact person: mleblanc@wheatoncollege.edu

Table of Contents

1. Background

Anglo-Saxon (used interchangeably with "Old English") was the language spoken in England from approximately 500 C.E. until 1066, when William the Conqueror imposed Norman French as the language of administration and law. Although Old English changed substantially in the wake of Scandinavian invasions in the ninth and tenth centuries, Anglo-Saxon texts from the seventh and eighth centuries were still intelligible to readers at the end of the period. During the century after the Conquest the spoken and written language changed to Middle English. Subsequent linguistic, cultural and historical change, including Henry VIII's dissolution of the monasteries (1536-1541) and later the English Civil War (1641-1653), combined to reduce radically the number of Anglo-Saxon texts, but some written Anglo-Saxon was saved. In the nineteenth century, techniques of *vergleichende Philologie* (the comparative and 'scientific' study of language change) developed by Rasmus Rask, Franz Bopp, Jakob Grimm, and others, enabled Anglo-Saxon texts to be read and understood for the first time in six hundred years. The long tradition of paleography, codicology and textual history based on comparative manuscript studies, originally developed for the study of Homer and Virgil but then applied to medieval manuscripts, unlocked hidden information about the relationships between texts and their connections to historical culture. Continued work in the twentieth century and into the present day shed new light on Anglo-Saxon culture. Computational and statistical approaches are the next step in this evolution.

2. Understanding Literature through 'Lexomics'

Originally coined for the field of bioinformatics, the term "lexomics" describes the computer-assisted detection and explanation of patterns, originally those in genomes (Dyer, Kahn and LeBlanc, 2007) but now those in any textual corpora. The lexomics approach seeks to identify subtle but distinctive word-use patterns across multiple texts, patterns that without computational and mathematical tools would be undetectable. Under the auspices of this NEH Digital Humanities Initiative start-up grant, we developed software tools and analytical methods that opened a new channel of information about relationships among Old English texts.

The specific relationships among many Old English texts have been vexed questions for nearly 200 years. Most Anglo-Saxon poetry is anonymous and exists only as tenth-century copies in manuscripts, though some of it is assumed to have been composed much earlier. We have only three named authors of poetry in the Anglo-Saxon period (Cædmon, Cynewulf and King Alfred), and there are various problems with linking these names with more than a very few specific poems. Although a number of prose texts are by known authors, even the majority of the prose is anonymous, or of questionable attribution. Thus, for years scholars of Old English have struggled to divine relationships between texts based on handwriting, script, physical characteristics of manuscripts, vocabulary, meter, style and the use of specific Latin sources. Computerized methods such as those developed under the auspices of this startup grant are easily applied across a wide range of texts, do not require serendipitous recognition of subtle patterns, and allow for discovery of additional interconnections among Anglo-Saxon texts. Information-processing acts like a microscope or telescope, allowing us to view things at a greater resolution that previously were not apparent. The methods developed—lexomic analysis, including clustering and classification according to statistical principles—allow us to extract meaningful patterns from complex data, to make new conjectures about relationships among texts and to highlight areas of interest to us, our collaborators and other scholars.

At one level, our lexomic analysis is very similar to what Gretsch and other researchers in the "Munich school" have done to discover specialized vocabulary in particular semantic contexts. We also look for word-use across different texts, but we have the capability of examining even more subtle patterns, sometimes even those in the frequency of very common words or "function words" (such as *and*/*ond*, *wiþ, þæt*), subtle but significant patterns that are unlikely to be detected by unaided human examination. Furthermore, our methods are automated allowing analyses of *many* texts, and sections of texts, in much less time than the more labor-intensive methods, providing researchers the ability to *screen* large textual corpora and to identify places where more detailed study is likely be most profitable.

3. Summary of Grant Activities

3.1 Who are we?

Our interdisciplinary group is comprised of an Anglo-Saxon scholar, a statistician, a computer scientist, three computer science undergraduate students, and a math undergraduate honors thesis student.

3.2 Initial NEH support and goals met

Beginning work in the summer of 2008 and continuing through the '08-09 and '09-10 academic years, well beyond the parameters of the Start-up Grant, the team met our original goals and more (see Table 1 for a summary):

(i) built a number of web-based tools for their group and other interested researchers (e.g., a "virtual manuscript" tool that allows researchers to mix and match poetry and prose text from multiple manuscripts into a single unit for investigation);

(ii) began a web-site for the dissemination of software and results;

(iii) implemented and tested a suite of open-source software;

(iv) made six conference presentations;

(v) submitted articles for peer-review with encouraging results that both support traditional scholarship and call for a renewed look at old questions; and

(vi) worked with scholars outside our group who were anxious to apply the lexomics software to their specific scholarly questions: Sarah Downey (California University of Pennsylvania), Yvette Kisor (Ramapo College), and Scott Kleinman (CalState Northridge).

(vii) purchased, installed, and continue to administer a high-powered server (Dual Core Xeon 1.86GHz with 32G of RAM, and an array of RAID disks for data backup) which hosts the group's website (**lexomics.wheatoncollege.edu**), disseminates our open-source software and documentation, and serves as an experimental workhorse for computational and memory-intensive experiments, for example, a cluster analysis on the majority of the corpus poetry.

(viii) offered our "connected" courses where we engage our undergraduates with the type of scholarship that we are doing in our research. Drout (English) teaches "Anglo-Saxon Literature" and LeBlanc (Computer Science) teaches "Computing for Poets". Four English majors have enrolled in an additional computer science course beyond the 'Poets' course; one math major is doing an honor theses.

Goal	Outcome	Notes
(1) (re)organization of *DOE* corpus to facilitate experiments	done (see scripts below)	
(2) suite of experimental software	done (see software below)	
(3) prototype website providing access to online tools to work with corpus and access to professionally documented open source software	done (see software and website below)	External Anglo-Saxon scholars have downloaded and experimented with our prototype software, applying it to their own queries; Set up Google Analytics software to monitor web-site usage
(4) strong connection of courses between English and Computer Science	Next iteration scheduled for Spring 2010 – Fall 2011	Recently, 4 English majors enrolled in a 2nd CS course beyond the connection. Two students are considering honor theses in this area.
(5) Identify consensus divisions within and relationships among Old English texts	Showed techniques worked on various Old English poems (*cf. JEGP* accepted paper and *Modern Philology* submission)	Discovered anomalies; research is leading to new publications
(6) External peer-review and dissemination	4 conference presentations 2 conference posters 1 newsletter article (accepted) 1 journal article (accepted) 1 journal article (submitted)	Our presentation at Congress 2009 led to three new collaborations with other Old English scholars

Table 1. Goals and outcomes of NEH Digital Humanities Start-up Grant ("Pattern Recognition through Computational Stylistics: Old English and Beyond", HD 50300-08, 2008-2010).

3.3 Results

Our results show the potential power of our hybridized lexomic and traditional techniques and are discussed in full in Drout, Kahn, and LeBlanc (2010, in press in *Journal of English and Germanic Philology*) and Drout *et al.* (under review in *Modern Philology*). For example, one portion of the long poem *Daniel* (found in MS Oxford, Bodleian Library, Junius XI) the "Prayer of Azarias" (lines 279-361), has a counterpart in a poem, *Azarias*, in a different Anglo-Saxon poetic codex, "The Exeter Book" (Exeter Cathedral Library 3501). *Azarias* corresponds roughly with the "Prayer of Azarias" in *Daniel*, but there are enough differences between the two versions for scholars to have concluded that neither poem was copied directly from the other and that instead they probably have some other source in common.

We were able to "recognize" that *Azarias* was more closely related to lines 279-361 of *Daniel* than to any other part of the poem (see dendogram of a cluster analysis on the next page). We first divided *Daniel* up into ten "chunks," each approximately the size of *Azarias,* and then used cluster analysis to determine which chunks of *Daniel* and *Azarias* were most closely related. The chunks of *Daniel* consisting of words 1801 through 2250, almost exactly the lines of *Daniel* that are paralleled in *Azarias,* are more closely related to *Azarias* than any other chunks of *Daniel*. This remarkable result indicates that the many small differences in words between *Daniel* and *Azarias* are 'coming out in the wash,' allowing us to extract and match more subtle patterns. Because these techniques examine hundreds of small points of comparison, and even though the program does not 'know' that *drihten* is the 'same' as *dryhten*, the method swiftly generates the correct answer. We are therefore more confident that, when comparing texts, our methods distinguish sections that are like each other. Even more significantly, these methods

work without laborious, subjective parsing or lemmatizing of texts. Even though the lexomic approach treats *cyninges* and *cyning* and *cyninge* as different "words" (rather than as inflected forms), these methods yield results consistent with the actual relationship of the texts. This eliminates the bottleneck in much computer-aided analysis—the marking up of texts by hand—and we can therefore examine very large groupings of texts, even the entire corpus.

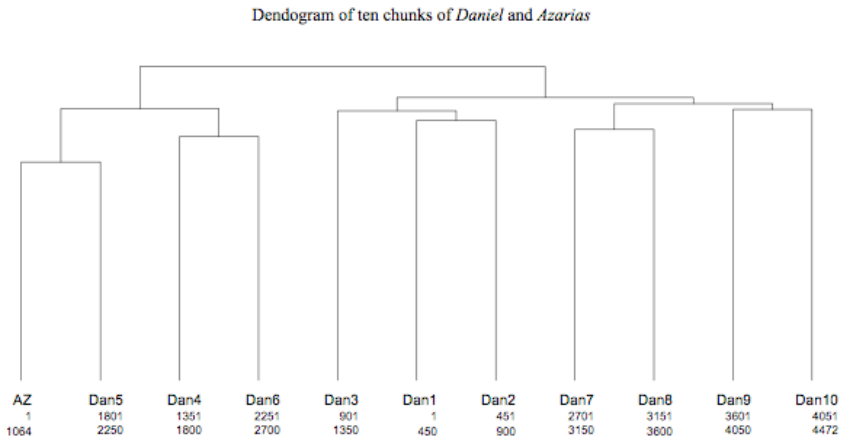


Figure 3. Dendogram showing the results of a cluster analysis using nine 450-word chunks of *Daniel* and one (ending) chunk of 422 words, and the entire 1064-word poem *Azarias* (AZ). The consecutive, non-overlapping 450-word chunks of *Daniel* are labeled from 1 to 10 where the initial chunk is labeled Dan1. The exact word boundaries of each chunk are labeled under each leaf on the dendogram, for example the fourth chunk of *Daniel* (Dan4) comprises the 450 words of *Daniel* from word 1351 to word 1800, inclusive.

Finding the section in one poem that is most like another poem did not exhaust the possibilities of our approach: we were also able to detect separate sections of a single long poem. The Old English *Genesis* is a verse paraphrase of the biblical story from the Creation to the sacrifice of Isaac. In 1875, the philologist Eduard Sievers noted that lines 235-851 of *Genesis* are significantly different in tone and style from the rest of the poem (lines 1-234 and 852-2936). He deduced that these lines, now called *Genesis B*, were a translation into Anglo-Saxon of an Old Saxon original, while the other lines of the poem, now called *Genesis A*, were a direct Anglo-Saxon translation of the Latin Vulgate. Sievers' deduction was confirmed when a fragment of an Old Saxon poem that matched some of the lines of *Genesis B* was found in the Vatican Library in 1894. The differences that allowed Sievers to recognize the source of *Genesis B* include variances in spelling, meter and style. Our software is not currently programmed to consider any of these characteristics *directly*, but using our lexomics approach, we looked at the differences in word-frequencies between various sections of *Genesis*. Again, we divided the poem into chunks, which we then tested for similarity with each other and, for the purpose of having outside comparanda, against chunks of other poems. Remarkably, our analysis accurately identified *Genesis B*, putting the three chunks that contain these lines separate from the rest of *Genesis*.

These early results function as *controls* whereby the known, basic relationships between *Azarias* and *Daniel*, as well as where *Genesis B* begins and ends are detected by lexomic analysis. To further extend this analysis, we examined the Exeter Book poem *Guthlac* to see if lexomic methods could detect separate sections in this poem. *Guthlac* is found on folios 32v-52v of the Exeter Book. Although the hand is the same throughout, at line 819 (folio 44v) a very large capital eth begins a line of large capitals, suggesting a significant division in the poem. In the dendogram below, *Guthlac B* is separated very clearly from *Guthlac A*, with B occupying all of clade ζ.

Dendogram of 1000 word chunks of *Guthlac A* and *B*

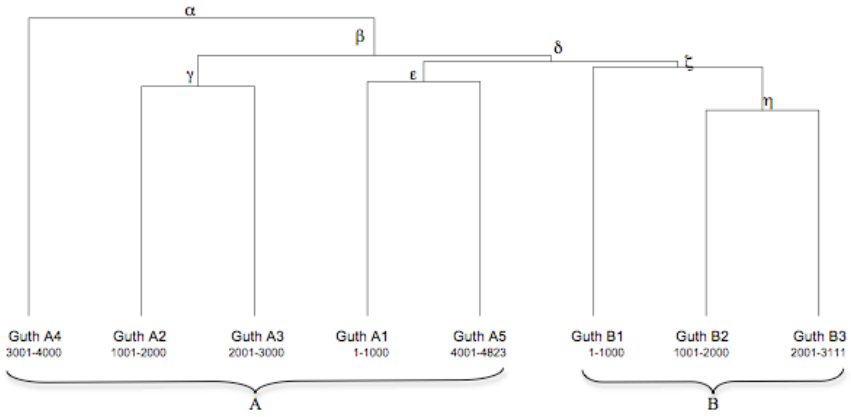


Figure 9. Dendogram showing the results of a cluster analysis using 5 approximately 1000 word chunks of *Guthlac A* and 3 approximately 1000 word chunks of *Guthlac B*.

The success of the lexomic approach in identifying the relationship of *Daniel* to *Azarias* and the divisions of *Genesis*, *Guthlac* and also of *Christ I,II, and III* suggests additional experiments that might shed light on the affinities of different poems. As argued in Drout *et al.* (2010, in press in *JEGP*), we conclude that the presence of a chunk in a simplicifolious clade indicates that the chunk in question is likely to have a different source than the source of the main body of the poem. This dendogram geometry, then, can be used to search for sections of poems that have outside sources, and it may shed some light on the composition methods of Anglo-Saxon poets.

Because our methods independently 'recognize' these known relationships, we are confident that these methods can identify other similarities and differences between and among poems. We propose to make further use of these methods to address hypothesized relationships and to discover unknown relationships, including those outside Old English. For example, we are particularly excited by analyses indicating that there should be an external source for a portion of *Genesis A* in which the poet augments the text of the Vulgate with a medieval tradition about the identity of Lamech. Our methods have told us where to look and what to look for, and we are

searching the *Patrilogia Latin* in an attempt to determine the exact Latin source of this section of *Genesis*.

3.4 Dissemination

Presentations:

May 2009

44th International Congress on Medieval Studies, Western Michigan University, Kalamazoo, MI.

(a) "Lexomics for Literature" Drout, M.C. and Kahn, M., Wheaton College

(b) Computing with Style: Investigations in Old and Middle English Poetry (A Panel Discussion).

Lexomics for Literature: Data
Michael Kahn, Wheaton College

Lexomics for Literature: Interpretation
Michael D. C. Drout, Wheaton College

(c) Poster Session sponsored by Digital Medievalist, the Medieval Academy of America Committee on Electronic Resources, and the Electronic Editions Advisory Board, Medieval Academy of America

July 2009

International Society of Anglo-Saxonists (ISAS 2009), Memorial University, St. John's, Newfoundland

"Ye shall have dominion over the sea [of texts]: Lexomics for Anglo-Saxon Literature"
Drout, M.C., Kahn, M., LeBlanc, M.D., and Nelson, C. '11

March 2010

The 41st ACM Technical Symposium on Computer Science Education (SIGCSE 2010), Milwaukee, WI

(a) "Connecting Across Campus" (LeBlanc, paper presentation)

(b) "Computing for Poets" (LeBlanc, poster presentation)

Papers:

LeBlanc, M.D., Armstrong, T., and Gousie, M. (March, 2010). "Connecting Across Campus". Published in the *Proceedings of 41st ACM Technical Symposium on Computer Science Education.*

Drout, M., Kahn, M., LeBlanc, M.D., Jones, A. '11, Kathok, N. '10, and Nelson, C. '11. "Lexomics for Anglo-Saxon Literature." *Old English Newsletter*, Spring 2010.

Drout, M., Kahn, M., LeBlanc, M.D. (accepted, to appear in 2010). Dendo-Grammar: Lexomic Methods for Analyzing Relationships Among Old English Poems. *Journal of English and Germanic Philology*.

Downey, S., Drout, M., Kahn, M., , LeBlanc, M. (submitted 2010) "Relationships Among Anglo-Saxon Guthlac Materials: A Confluence of Lexomic and Traditional Analyses." Under review at *Modern Philology.*

[in progress] Drout, M., Kahn M., LeBlanc M. (2010 in preparation). "Untangling Cynewulf: Lexomic Evidence for the Cynewulfian Canon."

[in progress] Salvador, M., Drout, M., Kahn, M., LeBlanc, M. (2011 in preparation). "The Compilation of the Riddles of the Exeter Book."

Papers continued …

[in progress] Drout, M., Kahn, M., Kisor, Y., LeBlanc, M. (2010 in preparation). "Lexomic Analysis of the Structure of *Beowulf*."

[in progress] Drout, M., Kahn M., LeBlanc M. (2010 in preparation). "The Devil Talks Like a Preacher Man: Lexomic and other Evidence for Borrowing from Homiletic Texts in the Speech of Demons and Devils in Anglo-Saxon Poetry."

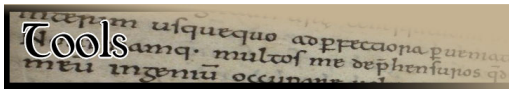
Publicity:

"Professors take new look at Old English texts." *Wheaton Quarterly*. Fall 2009.

Website: **`http://lexomics.wheatoncollege.edu`**

Online (browser-based) tools provide access to three online tools.

Stats by word: view the mean, minimum, median, maximum and standard deviation of each word across the entire Anglo-Saxon corpus. To enable further analysis on the researcher's own computer, the tools enable the user to download a zip-file of an Excel spreadsheet with the data.



Word Frequencies by: Text, Manuscript, Genre, or Corpus

View word frequencies. Use a drop-down menu to choose the level of detail for the statistics you desire. If you like the results, download a zipped Excel file to your Desktop for further analyses.

Build a Virtual Manuscript

We are most excited about our Virtual Manuscript prototype. As we obtained feedback from scholars using our tools, we learned that the organization in the online version of the *Dictionary of Old English* (DOE) may not facilitate many types of queries, for example, to mix and match poetry and prose texts from multiple manuscripts into a single unit for investigation. The "virtual manuscript" tool allows the user to pick and choose any number of texts, and the tool will generate the word frequency profile for that specific collection of texts. The user can download a zipped Excel file of the combined word counts and each selected text for further analyses.

Analysis Software

In addition to the software for the online tools (Perl and PHP), our suite of open source computational stylistic software (Perl) morphs data into needed formats in preparation for a researcher's experimental analyses of texts, including a pipeline of software used in our experiments to date. In addition, each script comes with a professionally documented README file (using Perl's pod format) to help with the use of each script.

as1_sort_Into_Directories.zip

Starting with a DOE directory of SGML-encoded texts, sorts texts into a new directory hierarchy according to the class/genre (A_Poetry vs. B_Prose) and manuscript names (A1, A2, ... B1, etc).

as2_cutter.zip
This handy script "cuts" texts into user-specified chunks. For example, you could cut the poem of *Daniel* into ten 450-word chunks; subsequent scripts will treat each of these chunks as an independent text.

as3_countWords.zip
This script counts the number of words in each file. Input files are organized in subfolders, e.g., by author or by text name if the text was split into chunks using a2_cutter.pl (see above). This workhorse script allows options for: (i) removing or keeping <tagged> words; (ii) consolidating common letters; (iii) sorting by words or by frequency; and (iv) disemvoweling all words or not.

as4a_mergeWordCounts.zip
This script can be used after you've created a Virtual Manuscript or following as3_countWords. The main goal of this script, in addition to collecting some statistics on your collection of texts, is to merge the counts into one file in preparation for further analysis, for example, in the statistical package, R (2008).

as4b_getStats_prepare4R.zip - ReadMe
This script is a follow-up to as4a. In short, this script generates additional statistical reports and rotates the merged data (rows to columns) for follow-up analyses in R.

Website Statistics of Use

In April 2009 we began using Google Analytics to monitor website use and traffic patterns. As of April 1, 2010, the site has received 342 (non-local) unique visitors from 30 countries with visitor loyalty of more than 10 visits for 43% of the visitors.

4. Recommended best practices

Multidisciplinary scholarship is hard … and that is good! (Not only for us as professors but for our undergraduate students as well). What really worked is that our culture of interdisciplinary teaching and research here at Wheaton made a seamless start to, and continuation of, this project. LeBlanc and Kahn's background of working in genomics provided the experience of "continual experiments", including many that do not work out.

We stress the value of closely linking research with teaching. In a cyclic fashion, our research spawns ideas to introduce in the classroom and then student work fuels new ideas for future scholarship. Drout and LeBlanc's connected courses (J.R.R Tolkien and Anglo-Saxon Literature connected with Computing for Poets) continues toward its fourth iteration and is widely recognized on campus as one of Wheaton's more successful connections.

This startup grant has energized the research side of this cycle, leading to a widening of our group. We cannot overestimate the value of including a statistician as part of the team. We now confidently show our students the value of interdisciplinarity *by example!*

This startup grant has become a model on our campus for how to start new interdisciplinary projects *and* get external funding for them. Our group did not anticipate how fun(!) this scholarship would be and we are all heartened by the encouragement and interest we are generating in English scholarship (both nationally at Kalamazoo, May 2009 and internationally at ISAS, July 2009) and the national computer science pedagogical conference (SIGCSE 2010).

We have a queue of external scholars who are either directly trying our tools, modifying our software, and/or seek to collaborate so we can design and implement experiments in their area of the corpus.

One of the keys to success was regular and frequent face-to-face meetings: we gathered all the PI's and most of the student collaborators nearly every Tuesday for two years. Treating the project this way, with a regular and consistent deadline of Tuesday morning, brought about much more rapid progress than other modes of meeting have in the past.

We also found that taking the time to teach each other was one of the most valuable ways we could spend our time. Teaching each other about Anglo-Saxon studies or statistics or computer science paid immediate dividends in creativity and problem solving; an hour spent teaching saved us many hours later through improved communication alone.

5. What We Might Have Done Differently

LeBlanc (PI, computer scientist) underestimated the amount of funding we'd need to keep everyone going for two full summers and for funding our dissemination plans. We lost a year of funding opportunity with our unsuccessful follow-up proposal to the NEH Collaborative Research area. In retrospect, we should have sought a wider range of counsel before applying to this program and/or participated as a grant reviewer at NEH.

We also should have, from the beginning, developed an organizational system and consistent nomenclature for our results. Although our dendograms are all saved, it would be helpful if the figures were stored in a central database that any of us could easily access (rather than digging through saved files). Naming conventions and consistency of format would also be very useful. The production of the first major research paper helped us a great deal, but we could still use more standardization and record management. Likewise, a better content management protocol would have helped us write the major papers a bit faster, though the great benefit of using a simple system like "track changes" in MS Word was helpful.

6. Next Steps

A significant result of this startup funding is the new set of collaborations that have emerged as a direct result of presenting on preliminary work in lexomics. We list the set of experiments that are currently being designed and tested with our new colleagues:

Guthlac and Anglo-Latin poetry with Sarah Downey. A long-standing problem in Anglo-Saxon studies is the relationship between a particular episode in the poem *Guthlac A* and various putative sources for the poem, the Latin *vita* of the saint, and the Old English translation of that vita and a variation of that translation found in the Vercelli Book. Our research shows that *Guthlac A* cannot have been the source for the Vercelli fragment but instead the fragment or, more likely, its source, was a source for the poem, further supporting the idea that the poem was written in the tenth century rather than the eighth. This work on the *Vita Guthlaci* by Felix has shown that we can apply analyze Latin texts as well as Anglo-Saxon texts and, with Prof. Downey, we are creating and analyzing a corpus of Anglo-Latin poems and prose texts. We have submitted our experimental results to the journal *Modern Philology* (February 2010).

Lemmatization of *Daniel* and *Azarias* with Scott Klienman (Cal State Northridge). Many colleagues raised the question of whether lexomics work when texts are lemmatized (that is, when each work is converted to its root form, so that "king" and "king's" are recognized as being the 'same' word). If lemmatized texts could be analyzed with the same success as the un-lemmatized texts we are currently using, our analysis could more easily cross the

linguistics boundary between Old English and Middle English or Middle English and Modern English. Prof. Kleinman has been preparing lemmatized versions of *Daniel*, *Azarias* and *Genesis*, as well as developing a set of mark-up conventions that will allow us to easily test both lemmatized and un-lemmatized texts.

Anglo-Saxon penitentials with Allen Frantzen (Loyola University Chicago). Prof. Frantzen has created a digital edition of the complete corpus of Anglo-Saxon penitentials that includes not only edited texts, but also diplomatic transcriptions of all manuscripts. We can therefore perform lexomic analysis on these manuscripts.

Beowulf with Yvette Kisor (Ramapo College). The most contentious problems in Anglo-Saxon studies have to do with *Beowulf*. Although lexomic analysis has not yet shed light on the biggest of these—the date of the poem—we can contribute to understanding of the structure and, perhaps, the composition of *Beowulf*. With Prof. Kisor we have divided up *Beowulf* into text units and then used lexomic methods to diagram the inter-relationship of these units. We are now using normalized and, eventually, lemmatized texts to further expand this analysis.

Riddles with Mercedes Salvador (University of Seville, Spain). We are working with Prof. Salvador, the leading expert on the Anglo-Saxon Riddles, to analyze the compilation of the Riddle collection in the Exeter Book. Lexomics sheds light on the relationships between different groups of Riddles and their sources. Preliminary work has already shown us that we can correctly identify the Riddles with immediate Latin sources in riddles by Aldhelm and Symphosius. Further work will help us to explain how the compiler of the Exeter Book expanded his original collection, perhaps in an effort to make the total number of riddles equal 100.

7. Participant Biographies

Dr. Mark D. LeBlanc (PI) – Co-Leader Wheaton Genomics Research Group and Professor of **Computer Science** at Wheaton College, Norton, MA.

Dr. LeBlanc is a co-leader of the Wheaton Genomics Research Group. His Ph.D. work in natural language processing involved the implementation of computational models to read and solve arithmetic word problems. Since 1998, LeBlanc has supervised the development of many software tools and computational experiments in genomics, including a web-based "DNA Dictionary" based on the online Oxford English Dictionary. In addition to his research in genomics, LeBlanc has received numerous grants from the National Science Foundation to develop and disseminate interdisciplinary course materials for bringing together faculty and students in biology and computer science. The most recent round of NSF funding resulted in a new textbook with Wheaton biology colleague Betsey Dyer: *Perl for Exploring DNA* (Oxford, 2007). Over the last two years in collaboration with Drout and Kahn and funded by this NEH Start-up Grant, LeBlanc has prototyped a suite of software to analyze the entire Old English Corpus. His undergraduate course entitled 'Computing for Poets' ("connected" with Drout's 'Anglo-Saxon Literature') teaches students in the Humanities to write computer programs in order to ask computational questions concerning large groups of texts and/or poems of interest.

Dr. Michael D.C. Drout (Co-PI) –Professor of **English** at Wheaton College, Norton, MA.

Michael D.C. Drout is Chair of the English Department and Prentice Professor of English at Wheaton College, Norton, MA, where he teaches Old and Middle English, fantasy and science fiction. Drout received a Millicent C. McIntosh Fellowship from the Woodrow Wilson Institute in 2006 and Wheaton's Faculty Appreciation Award for teaching in 2003.

Drout is the editor of J.R.R. Tolkien's *Beowulf and the Critics*, which won the Mythopoeic Scholarship Award for Inklings Studies for 2003, and *How Tradition Works: A Meme-Based Poetics of the Anglo-Saxon Tenth Century* (Arizona Medieval and Renaissance Studies 2006). He is one of the founding editors of the journal *Tolkien Studies* and is editor of the *J. R. R. Tolkien Encyclopedia: Scholarship and Critical Assessment* (Routledge 2006). The most-published Anglo-Saxonist of 2006, Drout has published on *Beowulf*, the Anglo-Saxon wills, the Old English translation of the *Rule of Chrodegang*, the Exeter Book 'wisdom poems' and Anglo-Saxon medical texts. Drout's English grammar book, *King Alfred's Grammar*, is available at his website, http://michaeldrout.com. His *Anglo-Saxon Aloud* daily podcast of the entire corpus of Old English poems has over 1500 regular listeners from 22 countries.

LeBlanc, Drout, Kahn

Dr. Michael J. Kahn (Co-PI) – Director of Quantitative Analysis and Professor of **Statistics** at Wheaton College, Norton, MA.

Dr. Michael Kahn has collaborated on numerous applications, mostly in the field of Biostatistics. His areas of research range from applications of Bayesian logistic regression models for health care funding, to analyses of cancer clinical trials and analyses of sociological data regarding parenting patterns in Jamaica. His most recent work is a collaborative effort regarding classification problems and techniques in genomics with Dr. LeBlanc and Dr. Betsey Dyer.